HHS Public Access

Author manuscript

Psychophysiology. Author manuscript; available in PMC 2022 June 01.

Published in final edited form as:

Psychophysiology. 2021 June ; 58(6): e13793. doi:10.1111/psyp.13793.

Standardized Measurement Error: A Universal Metric of Data Quality for Averaged Event-Related Potentials

Steven J. Luck[1,2], **Andrew X. Stewart**[1], **Aaron Matthew Simmons**[1], **Mijke Rhemtulla**[2]

[1]Center for Mind & Brain, University of California, Davis

[2]Department of Psychology, University of California, Davis

Abstract

Event-related potentials (ERPs) can be very noisy, and yet there is no widely accepted metric of ERP data quality. Here we propose a universal measure of data quality for ERP research—the *standardized measurement error (SME)*—which is a special case of the standard error of measurement. Whereas some existing metrics provide a generic quantification of the noise level, the SME quantifies the data quality (precision) for the specific amplitude or latency value being measured in a given study (e.g., the peak latency of the P3 wave). It can be applied to virtually any value that is derived from averaged ERP waveforms, making it a universal measure of data quality. In addition, the SME quantifies the data quality for each individual participant, making it possible to identify participants with low-quality data and "bad" channels. When appropriately aggregated across individuals, SME values can be used to quantify the combined impact of the single-trial EEG noise and the number of trials being averaged together on the effect size and statistical power in a given experiment. If SME values were regularly included in published papers, researchers could identify the recording and analysis procedures that produce the highest data quality, which could ultimately lead to increased effect sizes and greater replicability across the field.

Keywords

ERPs; reliability; replicability; signal-to-noise ratio; effect size; statistical power

1. INTRODUCTION

Event-related potentials (ERPs) are tiny signals, and they are embedded in noise that may be an order of magnitude larger. In theory, we can "average out" the noise by combining a large number of single-trial waveforms into an averaged ERP waveform. In practice, however, it is often difficult to obtain enough trials to adequately reduce the noise, and the remaining variability can dramatically reduce our power to detect significant differences. Moreover, the noise level may vary widely across recordings as a result of factors such as skin potentials, movement artifacts, poor electrode connections, and nearby electrical devices. The noise level may also be impacted by the experimental design, the recording procedure, and the

Contact information for corresponding author: Steven J. Luck, UC-Davis Center for Mind & Brain, 267 Cousteau Place, Davis, CA 95618, sjluck@ucdavis.edu, 530-297-4424 (phone), 530-297-4400 (fax).

signal processing pipeline. As a result, the signal-to-noise ratio may differ considerably across studies, across participants within a study, and across data processing methods.

1.1. Desirable properties for a metric of ERP data quality

Although noisy ERP waveforms are a major practical impediment in ERP research, the field has not adopted a universal measure of data quality that can be used to quantify the noise level in individual participants. Some metrics have been proposed, such as the root mean square of the voltage in the prestimulus period (Luck, 2014) or the standard deviation of a plus/minus average (Picton, 2011; Wong & Bickford, 1980). However, these metrics are insufficient because they are not designed to capture the noise that is *relevant* for a given experiment.

As an example, Figure 1 shows how high-frequency noise can distort the peak amplitude of the P3 wave (or any other component), adding considerable variability and decreasing statistical power. This high-frequency noise has much less impact when P3 amplitude is quantified as the mean voltage from 300–500 ms, because the upward and downward noise deflections largely cancel out (see Clayson, Baldwin, & Larson, 2013; and Chapter 9 in Luck, 2014). Thus, the impact of noise depends on the algorithm used to quantify a given amplitude or latency value, and one cannot make a generic quantification of the extent to which an EEG recording is “clean” or “noisy” independent of the method used to score the ERPs.

Traditional measures of ERP data quality do not take into account the role of the scoring method, and a central goal of the present work was to develop a metric of data quality that reflects the impact of noise on the specific amplitude or latency measure that is the focus of a given study. We will use the term *score* to refer to the actual amplitude or latency value being used as the dependent variable in a given experiment. When we use the phrase *data quality*, we are using *data* to refer to these scores (because they are the data that we are using to test our scientific hypotheses).

Some ERP studies have quantified the data quality for a specific score by using correlation-based measures of *reliability* (e.g., Boudewyn, Luck, Farrens, & Kappenman, 2018; Olvet & Hajcak, 2009; Thigpen, Kappenman, & Keil, 2017). However, this approach provides a single value for an entire group of participants rather than an individual data quality value for each participant[1]. Moreover, as described in detail in Section S1 of the online supplementary materials, correlation-based reliability measures are influenced by the range of true scores across individuals, so they do not provide a pure index of data quality.

We propose three criteria for a metric of data quality in ERP research: (a) the metric can be computed independently for each participant; (b) the metric quantifies the extent to which noise impacts the actual amplitude or latency scores that go into the statistical analyses and are used to test the scientific hypotheses; and (c) the metric represents the *precision* of the scores. Precision is defined as the degree to which repeated measurements under unchanged conditions yield similar results (Balazs, 2008). As discussed by Brandmaier et al. (2018) and

[1]A new reliability metric has recently been developed for individual participants (Clayson, Brush, & Hajcak, 2020).

illustrated in Figure 2, a metric of precision answers the question: If you were to obtain an ERP amplitude or latency score in a given participant multiple times in a given paradigm (assuming no learning, fatigue, etc.), how similar would the scores be across these repeated measurements? The importance of precision has been stressed more generally in statistical analyses by Vasishth et al. (2018).

As shown in Figure 2, a scoring method can be biased (shifted systematically in one direction away from the true value) independently of whether it is precise. This is typically a result of the signal processing methods or the algorithm used for scoring rather than the data per se, so we do not consider bias to fall under the umbrella of "data quality." For example, heavy low-pass filtering will bias onset latencies toward lower values (see Chapter 7 in Luck, 2014), and peak amplitude is biased to overestimate the true amplitude when high-frequency noise is present (see Chapter 9 in Luck, 2014). Thus, it is important that researchers consider whether the methods they are using to increase precision might also introduce or increase a bias.

1.2. Standardized measurement error (SME) as a universal measure of data quality for averaged ERPs

In this paper, we describe a simple but flexible metric of data quality for ERP research that meets these criteria. We call this metric the *standardized measurement error* or *SME*. The SME is a special case of the *standard error of measurement* that is designed to be well suited for ERPs. Specifically, the SME is defined as the standard error of measurement for an ERP amplitude or latency score, assuming that the score is obtained from a single participant's averaged ERP waveform. Formally, this means that the SME is an estimate of the standard deviation of the sampling distribution for a given participant's amplitude or latency score. Less formally, the SME indicates the extent to which the noise in the data has made an amplitude or latency score imprecise. The true SME is never known, but the following sections describe how it can be estimated, in which case we denote it as $\widehat{SME}$.

The SME can be used to quantify data quality for virtually any score that is obtained from an averaged ERP waveform, so it is a *universal* metric of data quality for averaged ERPs. However, it is specifically limited to scores obtained from averaged waveforms, and it is not designed for use with single-trial ERP analyses.

Widespread adoption of the SME would have many potential benefits for both individual researchers and the field as a whole. For example, individual researchers could use the SME to objectively and quantitatively determine whether data quality has been increased or decreased by a new recording procedure, signal processing method, or experimental design feature. If every ERP paper reported the SME, the field could objectively determine which recording and analysis procedures produce the cleanest data, and this would ultimately increase the number of true, replicable findings and decrease the number of false, unreplicable findings (Ioannidis, 2005; Vazire, 2018). Additional uses of the SME are described in Section 8.

To facilitate adoption of the SME, we have integrated it into version 8 of ERPLAB Toolbox (Lopez-Calderon & Luck, 2014; https://erpinfo.org/erplab), an open source Matlab package

for ERP processing and analysis. Because our code is open source, this metric of data quality could be easily added to other ERP analysis packages.

1.3. Structure of the paper

The remainder of this paper is structured as follows. Because the SME is a special application of the standard error of measurement, we begin by reviewing what a standard error actually quantifies and describing why it is ideal as a measure of data quality. Next, we describe how bootstrapping can be used to compute standard errors for virtually any amplitude or latency score that can be obtained from an averaged ERP waveform and how the SME allows us to define the signal-to-noise ratio in way that applies to even very complex scores (e.g., the onset time of a difference wave). Then we describe some practical considerations and assumptions involved in using the SME as a single-participant measure of data quality. Next, we describe how the SME can be aggregated across participants in a way that makes it possible to quantify the impact of data quality on the effect size and statistical power in a given study. Finally, we provide some preliminary recommendations about what constitutes a "good" SME value and how researchers could use the SME.

2. USING THE STANDARD ERROR AS METRIC OF DATA QUALITY

The SME is just a special application of the standard error of measurement. This section reviews what a standard error represents and then discusses why it is an ideal metric of data quality for averaged ERPs.

As an example of the general concept of a standard error, imagine a study in which we measured the height of 100 randomly selected women from the US and used the mean of this sample as an estimate of the mean of the entire population of US women. As illustrated in Figure 3A, we could estimate the standard error of this mean using an empirical approach. Specifically, we could repeat the study 10,000 times[2], with a different random sample each time, giving us 10,000 sample means. The distribution of sample means across these 10,000 samples would be an estimate of the *sampling distribution* of the mean, as shown in Figure 3A. We could then quantify the amount of variation among these sample means by taking the standard deviation ($\widehat{SD}$) of the 10,000 sample means. This $\widehat{SD}$ would be an estimate of the standard error of the sample mean (SEM)[3]. More generally, the standard error of a statistic is defined as the standard deviation of the sampling distribution for that statistic.

Now imagine that we had only 25 individuals in each sample. If we obtained 10,000 sample means with only 25 people per sample, as illustrated in Figure 3B, we would find much more variability across the sample means than we would find with samples of 100 people. The sampling distribution would therefore be broader, and the standard error would be larger. Thus, the standard error of a score reflects the precision[4] of the score (i.e., the extent

[2]In this and several subsequent examples, we consider what would happen with 10,000 replications of a study. There is nothing special about the number 10,000 in this context; it is just a convenient value for our examples because it is large enough to provide a very close approximation of what we would obtain with an infinite number of replications while being small enough to be within the bounds of imagination.

[3]Because we have a finite number of samples, this is only an estimate ($\widehat{SEM}$). To obtain the true standard error of the mean (*SEM*), we would need an infinite number of sample means.

to which we would expect the score to vary among different samples from the same population). In the analogy shown in Figure 2, each bullet hole is analogous to one sample mean, and the distribution of bullet holes is analogous to the sampling distribution. The sample means (bullet holes) form a tighter cluster when each individual sample mean is based on a larger number of samples[5].

Although the empirical approach for estimating the SEM is a useful mental exercise for explaining what a standard error represents, this approach is not typically used in practice because it requires obtaining many different samples to estimate the standard error. Fortunately, an analytic solution is available to estimate the SEM with the data from a single sample. Specifically, we simply compute the standard deviation ($\widehat{SD}$) of the single-participant scores in the sample and divide by the square root of the number of observations (N) in the sample:

$$\widehat{SEM} = \frac{\widehat{SD}}{\sqrt{N}} \quad \text{(Equation 1)}$$

The value we would get by applying Equation 1 to the data from a single study (the *analytic* $\widehat{SEM}$) is, on average, equivalent to the value we would get from the empirical method illustrated in Figure 3 (the *empirical* $\widehat{SEM}$). Section 3.2 will discuss a third way of estimating the SEM, called the *bootstrapped* $\widehat{SEM}$. These are just three different ways of estimating the same quantity. Importantly, none of these approaches assumes a normal distribution.

In addition, the samples can be trials rather than people, allowing us to examine the standard error of a single participant's mean score over trials. For example, when we construct an averaged ERP waveform for a single participant, we can think of this as the average of a set of single-trial EEG epochs that were sampled from an infinite hypothetical population of possible epochs for that one participant. In theory, we could repeat the recording session multiple times, each time obtaining a different sample of epochs and measuring an amplitude or latency score from the resulting averaged ERP waveform. The standard error of the score could then be estimated as the standard deviation of the scores obtained from the different recording sessions, and it would represent the precision of that score for that participant. This is illustrated in Figure 3C, which shows what would happen if we measured the peak amplitude of the P3 component from an ERP waveform created by averaging together 100 trials. If we repeated the recording session 10,000 times, obtaining a P3 amplitude from the averaged ERP waveform in each session, we could estimate the standard error as the standard deviation of those 10,000 values. Figure 3D shows that the sampling distribution becomes wider if only 25 trials are used to create the averaged ERP waveform, leading to a larger standard error.

[4]It might be more accurate to say that standard errors and the SME quantify the *imprecision* of a score, because they become larger as the score becomes less precise. However, it is more convenient to say that they "reflect" the precision of a score.

[5]In this example, the shots are centered on the bullseye (the true value). However, some measures are consistently shifted away from the true value, which is called *bias* in the measure. Here we are considering only the precision and not the bias of an amplitude or latency score.

In our height example, we took the *mean* height of a sample of 100 or 25 women, so our goal was to estimate the standard error of the *mean* (the SEM). In our P3 example, however, we measured the *peak* amplitude from a complex ERP waveform, which is not a mean. Thus, when we estimated the SD of the sampling distribution of our 10,000 repetitions, this value reflects the standard error of the peak amplitude rather than the standard error of the mean. As a result, we cannot estimate the standard error of this score using the data from a single session by using Equation 1.

Thus, we need a different method for estimating this standard error using the data from a single recording session. If we could accomplish this, the resulting standard error would meet the three criteria for a metric of ERP data quality described in Section 1.1: (a) it would be computed independently for each participant; (b) it would quantify the data quality for the specific scores that are entered into the statistical analyses; and (c) it would represent the precision of those scores. As the next section describes, the standard error can indeed be estimated for virtually any single-participant amplitude or latency score using the data from a single recording session. The standard error of a single-participant score can therefore serve as a universal metric of data quality for averaged ERPs. When the standard error of an ERP score is used in this way, we refer to it as the *standardized measurement error* (SME).

ERP waveforms are sometimes plotted with a shaded area showing the $\widehat{SEM}$ at each time point (see Supplementary Figure S1). As discussed in Section S2 of the online supplementary materials, this is not the same as the SME.

3. ESTIMATING THE SME

3.1. Estimating the SME for time-window mean amplitude scores

Because the SME is the standard error of measurement for a particular score, and many different types of scores are used in ERP research, it is important to select the appropriate method for estimating the SME for a given score. In this subsection, we describe a particularly common and straightforward case, in which the amplitude of an ERP component is scored from a single participant's averaged ERP waveform by calculating the mean voltage within a specific time window (e.g., 300–500 ms). We call this score the *time-window mean amplitude* (to distinguish it from a mean across trials or across participants). The next subsection will consider other kinds of amplitude and latency scores.

As an example of the time-window mean amplitude, we will consider data from a simple oddball experiment with 80 standard trials and 20 oddball trials (see Section S3 of the online supplementary materials for additional details). Figure 4 shows the single-trial EEG epochs and the averaged ERP waveforms from a single participant in this experiment. In this example experiment, we scored the amplitude of the P3 wave as the time-window mean amplitude from 300–500 ms, measured from the averaged ERP waveforms separately for each participant.

To estimate the SME for this score for a single participant, we would simply obtain one score from each single-trial EEG epoch and apply Equation 1 to these single-trial scores. That is, we would compute the time-window mean amplitude from 300–500 ms on each

individual trial, calculate the $\widehat{SD}$ of these scores, and divide by the square root of the number of trials. We would do this separately for the oddball and standard trials. For the participant shown in Figure 4, we obtained an $\widehat{SD}$ of 4.96 μV and an $\widehat{SME}$ of 0.56 μV for the standards and an $\widehat{SD}$ of 5.73 μV and an $\widehat{SME}$ of 1.28 μV for the oddballs. Note that the $\widehat{SD}$ values were similar for the standards and the oddballs, but because Equation 1 requires us to divide the $\widehat{SD}$ by the square root of the number of trials ($\sqrt{80}$ for the standards and $\sqrt{20}$ for the oddballs), the $\widehat{SME}$ was much larger for the oddballs than for the standards.

In this example, the precision of our estimate of the time-window mean amplitude was better (i.e., the $\widehat{SME}$ was lower) for the condition with more trials, as one would ordinarily expect. Note that the data quality of the single-trial EEG was approximately the same for the standards and for the oddballs (i.e., the $\widehat{SD}$ of the single-trial values were similar), but the data quality of the time-window mean amplitude scores obtained from the averaged ERPs was much better for the standard than for the oddballs (because of the difference in the number of trials in each average). This illustrates one of the fundamental goals of the SME, which is to provide a metric of data quality for the actual amplitude and latency scores that we obtain from averaged ERP waveforms and use as the dependent variables in our statistical analyses. The SME is sensitive to any factor that produces imprecision in these scores, including the number of trials in each average and the amount and type of noise in the single-trial EEG epochs.

3.2. Extension to other ERP amplitude and latency scores with bootstrapping

Equation 1 can be applied to the time-window mean amplitude score because this score has the same value whether we obtain it from the averaged ERP waveform (as is usually done) or by measuring it on each individual trial and then averaging those single-trial scores together. This is illustrated in Figure 5, which shows time-window mean amplitude scores obtained from a set of single trials and from the average of those trials. Because we get exactly the same value for the time-window mean amplitude whether we score the averaged ERP waveform or score the single-trial epochs and then average the single-trial scores together, the standard error of the mean of the single-trial scores is also the standard error of the score obtained from the averaged ERP waveform. That is, the two different ways of computing the time-window mean amplitude score always produce exactly the same value, so they have the same standard error.

Unfortunately, Equation 1 cannot be validly applied to most other amplitude or latency scores that are obtained from averaged ERP waveforms (e.g., peak amplitude or peak latency). For example, consider what would happen if we used the peak amplitude between 300 and 500 ms as our score. If we measured the peak amplitude scores from the individual trials and then averaged those scores together, the result would not be the same as measuring the peak amplitude from the averaged ERP waveform (see Figure 5). Because the average of the single-trial scores is not the same as the score obtained from the averaged ERP waveform, the standard error obtained from the mean of the single-trial peak amplitude scores is not the same as the standard error of the peak amplitude score obtained from the averaged ERP waveform. They are different scores, so they have different standard errors. Thus, Equation 1 cannot be used for peak amplitude.

The same issue applies to virtually every ERP amplitude or latency measure other than the time-window mean amplitude (including peak amplitude, peak latency, and every widely used measure of onset or offset latency[6]). Could we then just use the mean of the single-trial scores as our dependent variable (rather than obtaining the scores from the averaged ERP waveforms), so that we could use Equation 1 to estimate the standard error? Unfortunately, the answer is usually "no" for scores other than the time-window mean amplitude[7], because these other scores are typically quite distorted by the single-trial noise. As a result, our statistical power would be dramatically reduced by using the average of the single-trial scores as the dependent variable. Consequently, scores other than the time-window mean amplitude are ordinarily obtained from averaged ERP waveforms, and we need a different method for estimating the standard errors for these scores.

Fortunately, we can use *bootstrapping* to estimate the standard error for the peak amplitude, the peak latency, or virtually any score that can be obtained from an averaged ERP waveform[8]. The logic behind bootstrapping is described in detail in Section S4 of the online supplementary materials, but we provide a brief overview and example here.

Bootstrapping provides a means of simulating the empirical procedure for estimating a standard error (see Figure 3B), in which we obtain the score in a large number of experiments and use the $\widehat{SD}$ of these scores as our estimate of the standard error. Instead of repeating the experiment many times, we can simulate new experiments from an existing set of n trials[9] by randomly sampling n trials from the existing data set *with replacement*. For example, if we have 20 trials in a given experiment, we would simulate a new experiment by sampling 20 trials at random *with replacement* from this set of trials and then averaging over those 20 trials. Because this procedure involves sampling with replacement, we get a different set of n trials and therefore a different averaged ERP waveform every time we simulate an experiment (assuming that n is sufficiently large). We can obtain our amplitude or latency score from the averaged waveform for each simulated experiment. This gives us one score for each simulated repetition of the experiment, and the $\widehat{SME}$ is the $\widehat{SD}$ of these scores (because the standard error is the SD of the sampling distribution). In other words, rather than having to repeat the recording session many times with each participant, we can simulate a large number of sessions by sampling with replacement, giving us an estimate of the sampling distribution so that we can compute the $\widehat{SD}$ of this distribution. Moreover, this approach focuses on what a standard error actually represents, namely the SD of a sampling distribution[10].

[6]More specifically, Equation 1 can be used only for scores that are computed by linear transformations of the ERP data (see the Appendix in Luck, 2014).

[7]Although it is rare, these scores are sometimes obtained from single-trial EEG epochs. In these cases, Equation 1 could be used to estimate the SME. However, the resulting SME value would reflect the precision of the mean of the single-trial scores, not the precision of the score that would be obtained from the averaged ERP waveform.

[8]It is possible that analytic solutions like Equation 1 will someday be developed (or may already exist) for estimating the standard error for these other scores. If that happens, these analytic solutions could potentially be used instead of bootstrapping. However, bootstrapping requires minimal assumptions, and it can be applied to the result of an arbitrary sequence of data processing operations (e.g., the onset latency of a low-pass filtered difference wave). Thus, we emphasize bootstrapping as the best practical solution in most cases at the present time.

[9]We are using n to refer to the number of trials and N to refer to the number of participants. Either n or N can be used in the denominator of Equation 1.

As an example, we used bootstrapping to compute the $\widehat{SME}$ for the peak latency of the P3 wave for oddballs and standards for the same participant shown in Figure 4B. Figure 6 shows the data from this participant. The single-trial EEG epochs for the standards and the oddballs are shown in Figure 4B, and the conventional averaged ERP waveforms (i.e., the averages of the 80 standards and the 20 oddballs) are shown in Figure 6A. Panels B and C of Figure 6 show averaged ERP waveforms for two iterations of the bootstrapping procedure. On each iteration, 80 trials were selected at random *with replacement* from the set of 80 standards, and 20 trials were selected at random *with replacement* from the set of 20 oddballs. You can see that the resulting waveforms from these bootstrap iterations are similar but not quite identical to the conventional averaged ERP waveforms. Figure 6 also shows the peak latencies scored from these waveforms (the time of the maximum voltage between 300 and 500 ms).

We iterated this process a total of 10,000 times, each time creating new averaged ERP waveforms for the standards and for the oddballs from randomly selected sets of trials and then measuring the peak latencies from these new waveforms. This gave us 10,000 peak latency scores for both the standards and the oddballs. The $\widehat{SME}$ for a given trial type is simply the $\widehat{SD}$ of these 10,000 scores. For the standards, the peak latency from the conventional averaged ERP waveform was 345 ms, and we obtained an $\widehat{SME}$ of 35.2 ms. For the oddballs, the peak latency was 459 ms, and we obtained an $\widehat{SME}$ of 16.6 ms.

Note that the peak latency SME was actually worse (larger) for the standards than for the oddballs, even though there were many more trials for the standards than for the oddballs. By contrast, the SME for mean amplitude was approximately half as large for the standards as for the oddballs (see Section 3.1). The relatively large SME for peak latency in the standards is a result of the fact that the waveform for the standards was relatively flat in this participant, without a clear peak. As a result, noise in the data can cause fairly large variations in the time point at which the waveform reaches its maximum value. The oddball waveform had a much clearer peak, so noise had less impact on the latency score. Indeed, if you compare the peak latencies in Figure 6 for the two bootstrap iterations, you'll see that the latencies are within 10 milliseconds for the two oddball waveforms but are nearly 100 ms apart for the two standard waveforms. This is exactly what we would expect for this participant if we actually repeated the experiment multiple times: because the waveform for the standards has no clear peak, we would expect more variation in the peak latency for the standards than for the oddballs. This demonstrates how the effect of noise for a score like peak latency may depend on complex factors such as the shape of the waveform, making our bootstrap-based approach particularly valuable for such scores.

3.3. Additional SME examples

The example shown in Figure 6 illustrates how we obtained the bootstrapped $\widehat{SME}$ for the peak latency in a single participant. We also obtained bootstrapped $\widehat{SME}$ values for the other

[10]Equation 1 and the bootstrapping method described here are two of many possible ways of estimating the standard error of measurement, and we are not wedded to these particular estimation approaches. The general logic of the SME is independent of the method used to estimate the standard error.

11 participants in this study; the Matlab code, data, and results are provided at https://doi.org/10.18115/D58G91. We also provide code, data, and results for the time-window mean amplitude from 300–500 ms (estimated using Equation 1 and also using bootstrapping) and for peak amplitude for the 300–500 ms time window (estimated using bootstrapping). The single-participant $\widehat{SME}$ values for each of these scores are provided in Table 1, along with the actual amplitude and latency scores. Additional details, figures, and discussion are provided in Section S5 of the online supplementary materials.

Table 1 shows that the analytic and bootstrapped values for the time-window mean amplitude are typically very similar. Because the analytic SME is trivial to compute for the time-window mean amplitude score, ERPLAB Toolbox automatically computes the analytic SME whenever an averaged ERP is computed, using either default or custom time windows. The bootstrapped SME is more complicated to compute and currently requires simple Matlab scripting.

The bootstrapped SME can also be computed for scores that require multiple processing steps after averaging. For example, to determine whether an experimental manipulation influences the starting time of a cognitive process, it can be useful to measure the onset latency of a difference wave that isolates that process (Luck, 2014). This is particularly common in experiments that use the lateralized readiness potential to assess the onset of motor activation and experiments that use the N2pc component to assess the time at which visual attention has shifted to a target (Luck, 2012; Smulders & Miller, 2012). These components are isolated from the rest of the ERP waveform by means of a contralateral-minus-ipsilateral difference wave, and the onset latency is typically measured from this difference wave. Bootstrapping can be used to estimate the SME for these onset latency measures by simply creating a contralateral-minus-ipsilateral difference wave on each iteration and then measuring the onset latency from that difference wave. To demonstrate this general approach, the code provided at https://doi.org/10.18115/D58G91 includes an example in which the peak latency of the P3 wave is measured from a rare-minus-frequent difference wave.

It may also be useful to score the time-window mean amplitude from a difference wave. The corresponding SME will quantify the precision of the difference in amplitude between conditions (see Thigpen et al., 2017 for an analogous application of Cronbach's alpha). Other operations that might be performed prior to scoring an amplitude or latency would include averaging multiple electrodes into a cluster, filtering the waveforms, and application of the Laplacian transform.

4. INTERPRETATION OF THE SME

To reiterate, the SME is a metric of precision and therefore answers the question: If we repeated the experiment an infinite number of times with a given participant (with no learning, fatigue, etc.), and we obtained the participant's amplitude or latency score on each repetition, how variable would those scores be? More precisely, the SME allows us to estimate the standard deviation of the values that would be obtained across repetitions of the

experiment. The units of the SME are on the same scale as the score (e.g., μV for a typical amplitude measure, ms for a typical latency measure).

Any source of trial-to-trial variability that influences a participant's amplitude or latency score is considered a source of error. This includes induced electrical activity from the recording environment (e.g., line noise), biological artifacts (e.g., skin potentials), movement artifacts, EEG signals that are not phase-locked to the time-locking point (e.g., alpha-band EEG oscillations), and trial-to-trial variation in the amplitudes or latencies of the underlying ERP components. For example, alpha-band EEG oscillations can add substantial trial-to-trial variability to peak amplitude scores, so these oscillations are considered a source of measurement error with respect to these scores. In a different study, however, these same alpha-band oscillations could be the signal of interest rather than a source of measurement error. Similarly, trial-to-trial variability in the ERP component of interest might be of considerable theoretical interest, but it would be considered a source of measurement error if the component is scored from averaged ERP waveforms. Thus, when researchers obtain scores from averaged ERP waveforms (as in the vast majority of current ERP research), the SME reflects all sources of imprecision for that score (e.g., biological and nonbiological artifacts, mind wandering, lapses of attention). Consequently, the SME would not be suitable for determining whether neural variability varies across individuals, groups, or experimental conditions. Moreover, the SME is a metric of data quality for scores obtained from averaged ERP waveforms, and it was not designed for use in studies that focus on single-trial data.

Note that SME will be influenced by any signal processing operations that have been applied to the data before the score of interest has been obtained (e.g., filtering, re-referencing, artifact rejection). This is exactly what we want for a metric that quantifies the precision of the scores that are entered into our final statistical analyses. Indeed, the SME will be useful for determining whether a given signal processing operation increases or decreases the precision of the scores. The SME will also depend on the number of trials, which is again exactly what we want for a metric that quantifies the precision of our actual scores.

The SME can also be used to estimate the signal-to-noise ratio of a given ERP score. The score is our estimate of the signal, and the SME is our estimate of the noise for that score, so our estimate of the signal-to-noise ratio is simply the score divided by the SME. Additional details are provided in Section S6 of the online supplementary materials.

5. PRACTICAL ISSUES, ASSUMPTIONS, AND CAVEATS

Using the SME to quantify data quality requires some assumptions and raises some practical issues, which we will address in this section.

As a practical matter, we must deal with the fact that some trials may be rejected because of artifacts, leading to different numbers of trials per participant or per condition. The SME naturally takes the actual number of trials in each averaged ERP into account (e.g., as the denominator in Equation 1), so the SME values reflect the quality of the data that results from the actual number of trials in the averaged ERP waveform. Similarly, if the

experimental design yields more trials in some conditions than in others, the SME will be better (smaller) for the conditions with more trials (all else being equal).

You might wonder about the minimum number trials needed to use bootstrapping. If too few trials are available, you may sample exactly the same set of trials on different iterations. However, this is not typically a problem in practice if you have at least 8 trials (Chernick, 2011). Also, you should keep in mind that both bootstrapping and Equation 1 merely provide an *estimate* of the standard error, and the precision of this estimate will be greater if you have more trials. In other words, both the data quality and your ability to accurately estimate the data quality may be poor when the number of trials is small[11]. It is difficult to quantify the number of trials needed for an acceptable estimate, but a practical approach is to determine whether multiple repetitions of the bootstrapping procedure converge on similar SME values (Chernick, 2011). We have found reasonably good convergence with 20 trials, but simulations with a broad set of experimental conditions are needed before firm recommendations can be provided.

Both Equation 1 and bootstrapping assume that the individual trials are independent of each other, but this assumption will typically be violated because the trials are actually obtained sequentially from a brain that may gradually change state over time and that adapts in response to experience. Section S7 of the online supplementary materials explains why violations of this assumption are unlikely to be a major problem in most cases and describes how the SME estimation procedure could be modified if sequential dependencies turn out to be problematic.

6. AGGREGATING ACROSS PARTICIPANTS TO PREDICT TRUE SCORE VARIANCE, EFFECT SIZES, AND STATISTICAL POWER

Up to this point, we have focused on using the SME to quantify data quality for individual participants, and we have shown that it meets the three criteria specified in Section 1.1. However, the SME has another virtue when it is aggregated across participants in a specific manner: the aggregated group SME can be used to estimate the portion of the total variance (Var_{Total}) across individuals that reflects measurement error and the portion that reflects true differences among individuals (much like traditional psychometric reliability measures). Using this information, you could determine how increasing or decreasing the number of trials would impact your effect size and statistical power (i.e., the probability of obtaining a significant effect if, in fact, a real effect exists). This aggregated SME value can also tell you how well the data quality in one experiment compares to the data quality in another experiment and how well your lab's data quality compares with the data quality of other labs. Conventional measures of variability (e.g., the group $\widehat{SD}$) are insufficient to achieve these goals, because they are influenced both by measurement error (which can be modified by changing the number of trials, the filter settings, etc.) and true differences among individual participants.

[11]Both the analytic and bootstrapped SME estimates tend to underestimate the true standard error when the number of trials is small. However, the degree of underestimation is modest and should have little or no practical impact for the most common uses of the SME (as described in Section 8).

6.1. Aggregating across participants with RMS($\widehat{SME}$)

It would be possible to summarize the data quality from a given experiment with the mean of the single-participant $\widehat{SME}$ values. However, there is a better approach that links the SME values directly to the effect size and statistical power of a given experiment. In this approach, the $\widehat{SME}$ values obtained for the individual participants in a group are combined using the *root mean square* (RMS) of the values. We call the resulting value $RMS(\widehat{SME})$. When applied to the set of $\widehat{SME}$ values obtained from each member of a group of N participants ($\widehat{SME}_{1:N}$), $RMS(\widehat{SME})$ is defined[12] as:

$$RMS\left(\widehat{SME}_{1:N}\right) = \sqrt{\frac{\widehat{SME}_1^2 + \widehat{SME}_2^2 + \widehat{SME}_3^2 + \cdots + \widehat{SME}_N^2}{N}} \quad \text{(Equation 2)}$$

We recommend reporting $RMS(\widehat{SME})$ to summarize the data quality from a given condition of an experiment. Like single-participant $\widehat{SME}$ values, RMS($\widehat{SME}$) is in the same units as the score (e.g., μV for most amplitude measures, ms for most latency measures). This makes it easy to compare with the actual scores (e.g., an RMS($\widehat{SME}$) of 40 ms relative to a mean P3 peak latency of 426 ms). However, equations relating the $\widehat{SME}$ to effect size and statistical power are simpler if you square the RMS($\widehat{SME}$) value (or just don't take the square root when computing the RMS($\widehat{SME}$)). This would then be the *mean square* of the $\widehat{SME}$ or $MS(\widehat{SME})$. That is, if you have N participants,

$$\begin{aligned} MS\left(\widehat{SME}_{1:N}\right) &= RMS\left(\widehat{SME}_{1:N}\right)^2 \\ &= \frac{\widehat{SME}_1^2 + \widehat{SME}_2^2 + \widehat{SME}_3^2 + \cdots + \widehat{SME}_N^2}{N} \end{aligned} \quad \text{(Equation 3)}$$

These values are no longer in the same units as the score itself (e.g., if $RMS(\widehat{SME})$ = 40 ms for a P3 peak latency score of 426 ms, then $MS(\widehat{SME})$ = 1600 ms^2), making it a less natural descriptive statistic than RMS($\widehat{SME}$). However, $MS(\widehat{SME})$ is more convenient for some of the equations found in the next section.

6.2. Decomposing variability across participants into true score variance and measurement error

We now turn to the relationships among $MS(\widehat{SME})$, number of trials, effect size, and statistical power. First, however, it necessary to briefly review how measurement error is treated by classical measurement theory. Imagine that you've conducted an ERP experiment, and you have measured P3 latency from each participant's averaged ERP waveform. You

[12] Some of the notation used in our equations is different from the conventions of the statistics literature. Our goal is to ensure that the equations are easily understood by individuals without a background in mathematical statistics, even if this makes the equations somewhat less compact and nonstandard. Anyone with a strong background in statistics should be able to translate these equations into more conventional formats. However, we do follow the convention of using an italicized variable name to represent a parameter of a population and the same name with a caret over it to represent an estimate of that parameter (e.g., SD for the standard deviation of a population and $\widehat{SD}$ for the estimated standard deviation from a sample of this population).

would calculate the sample mean of these scores across participants, and you could also calculate the estimated variance ($\widehat{Var}$) or the $\widehat{SD}$ of the scores to quantify the variability across participants. This variability reflects a combination of two factors: true differences among individuals (i.e., differences that would be present even if we had an infinite number of trials per participant) and measurement error (i.e., differences among individuals that are a result of trial-to-trial variability in the data rather than stable differences between people[13]).

We call the total variance in scores[14] across individuals Var_{Total} (which is the same as the square of the total SD, SD_{Total}). We call the variance that is caused by true differences across individuals the *true score variance* (Var_{True}), and we call the variance that is caused by measurement error the *measurement variance* ($Var_{Measurement}$). Because variances simply sum together (for independent random variables), we can express the total variance as the sum of the true score variance and the measurement variance:

$$Var_{Total} = Var_{True} + Var_{Measurement} \quad \text{(Equation 4)}$$

Because single-participant SME values quantify measurement error, MS($\widehat{SME}$) can be used to estimate $Var_{Measurement}$, and we can therefore rephrase Equation 4 in terms of estimates of the various terms:

$$\widehat{Var}_{Total} = \widehat{Var}_{True} + MS\left(\widehat{SME}\right) \quad \text{(Equation 5)}$$

We can then estimate the true score variance by simply rearranging the terms of Equation 5 as:

$$\widehat{Var}_{True} = \widehat{Var}_{Total} - MS\left(\widehat{SME}\right) \quad \text{(Equation 6)}$$

We can convert $\widehat{Var}_{True}$ back into SD units by simply taking the square root:

$$\widehat{SD}_{True} = \sqrt{\widehat{Var}_{True}} = \sqrt{\widehat{Var}_{Total} - MS\left(\widehat{SME}\right)} \quad \text{(Equation 7)}$$

In other words, we can now quantify how much of the variability across participants in our score ($\widehat{SD}_{Total}$ or $\widehat{Var}_{Total}$) is a result of real differences among people ($\widehat{SD}_{True}$ or $\widehat{Var}_{True}$) and how much is a result of measurement error ($\widehat{SD}_{Measurement}$ or $Var_{Measurement}$)[15].

[13]For the sake of simplicity, we are not taking into account variations in scores that are a result of variations in the testing conditions or the state of the participant.

[14]We would like to reiterate that we are using the term *score* to refer to an amplitude or latency value obtained from the averaged ERP waveform from a single participant.

[15]The amount of variability produced by participants and measurement error is more easily expressed in units of variance than units of SD. For example, $Var_{Measurement}$/ Var_{Total} can be used in a straightforward way to quantify the proportion of variance due to measurement error. However, SD values do not combine in an additive manner, so it would not be appropriate to use SD_{ME}/ SD_{Total} to quantify the proportion of the total SD that is explained by measurement error.

This also makes it possible to quantify the *psychometric reliability* of our scores, where psychometric reliability is defined as the proportion of total variance that is accounted for by true score variance[16]:

$$Reliability = \frac{Var_{True}}{Var_{Total}}$$

This can be estimated as:

$$\widehat{Reliability} = \frac{\widehat{Var}_{Total} - MS\left(\widehat{SME}\right)}{\widehat{Var}_{Total}} = 1 - \frac{MS\left(\widehat{SME}\right)}{\widehat{Var}_{Total}} \quad \text{(Equation 8)}$$

Thus, the SME has the advantage of being computed for individual participants but can also be converted into group-level psychometric reliability (yielding a value that would be similar to that produced by traditional approaches, such as split-half reliability). Note, however, that we recommend against using group-level psychometric reliability as a general metric of ERP data quality (see Section S1 of the online supplementary materials).

6.3. *MS*($\widehat{SME}$), effect size, and statistical power

We will now briefly describe how $MS(\widehat{SME})$ makes it possible to predict effect sizes and statistical power. In a simple experiment with two groups, the effect size (Cohen's *d*) is defined as the difference in means between the two groups divided by the pooled SD_{Total} values for the two groups. The effect size therefore depends directly on the SD_{Total} values. Equations 4–7 show how SD_{Total} is related to $MS(\widehat{SME})$, and we can use these equations to predict how SD will vary as we increase or decrease the measurement error.

We can then ask, for example, how the effect size would vary if we increased or decreased the measurement error (e.g., by increasing the number of trials or by decreasing the single-trial EEG noise). Further, because statistical power is related to effect size, Equations 4–7 make it possible to estimate how power will change if the number of trials is increased or decreased (under the assumption that the $\widehat{SME}$ values will be proportional to the square root of the number of trials[17]). Indeed, Baker et al. (2020) have provided a power calculator that allows you to use the measurement error variance (which can be estimated by $MS(\widehat{SME})$) and the true score variance (which can be estimated using Equation 6) to estimate how your statistical power will change as a joint function of the number of trials and the number of participants.

[16]Note that this is how reliability is typically defined in psychometrics, but the term *reliability* is defined quite differently in fields such as physics and engineering (see Brandmaier et al., 2018). We therefore use the term *psychometric reliability* to be clear that we are discussing this particular meaning of the term.

[17]Note that this "square root rule" is only an approximation. First, the noise level in the data may increase or decrease as the length of the experiment changes (due to factors such as learning and fatigue). Second, for scores other than the time-window mean amplitude, the standard error may not be linearly related to the square root of the number of trials. Third, even if the single-participant $\widehat{SME}$ values change linearly with the number of trials, the effect of the number of trials on the aggregated $RMS(\widehat{SME})$ value will depend on the distribution of single-participant SME values. However, it would still be reasonable to use this assumption to make educated guesses about the impact of changes in the number of trials.

The use of $MS(\widehat{SME})$ to see the relationship between data quality and statistical power assumes that the amplitude or latency scores will be obtained from averaged ERP waveforms. However, time-window mean amplitude scores can be obtained just as well from the single-trial EEG epochs (see Section 3.1). The single-trial scores can then be analyzed using multilevel models, which have many advantages over conventional *t* tests and ANOVAs (Bürki, Frossard, & Renaud, 2018; Volpert-Esmond, Merkle, Levsen, Ito, & Bartholow, 2018; Winsler, Midgley, Grainger, & Holcomb, 2018). In this single-trial analysis approach, the $\widehat{SME}$ no longer reflects the precision of the dependent variable used in the statistical analysis, and $MS(\widehat{SME})$ is not directly related to the effect size and statistical power. However, this approach is not yet widely used in ERP research, and scores other than the time-window mean amplitude must usually be obtained from averaged ERP waveforms. Thus, the $\widehat{SME}$ is well suited for most current ERP research. As the field moves toward multilevel models of single-trial data, it may be possible to quantify measurement error as the standard error of the single-subject slope or intercept values (see Bürki, Elbuy, Madec, & Vasishth, 2020).

7. WHAT IS A GOOD OR BAD $\widehat{SME}$ VALUE?

We now consider the question of what would be considered a "good" or "bad" $\widehat{SME}$ value in an actual experiment. This is not a simple question, and it will likely depend on the experimental design, the participant population, and the scientific hypothesis being tested. If $\widehat{SME}$ values become widely reported, it will be possible to see what values are typical in various types of experiments, and these typical values will serve as anchor points for determining what values are "good" and "bad" in practice. In addition, individual laboratories can calculate $\widehat{SME}$ values for previous experiments to see the range that is typical in their experiments. In the meantime, we provide some preliminary heuristics for defining "good" and "bad" $\widehat{SME}$ values.

The first heuristic is to compare the RMS($\widehat{SME}$) values to the $\widehat{SD}$ of the observed scores across participants ($\widehat{SD}_{Total}$). If RMS($\widehat{SME}$) is much smaller than $\widehat{SD}_{Total}$, this would indicate that the observed differences across individual participants are mainly driven by true individual differences, with relatively little impact of measurement error. By contrast, if the RMS($\widehat{SME}$) is close to the $\widehat{SD}_{Total}$ value, this indicates that most of the observed variability is a consequence of measurement error. Examples are provided in Section S5 of the online supplementary materials.

It may be useful to exclude participants with extreme $\widehat{SME}$ values, because the measured scores for those participants are likely to be poor estimates of their true scores. There are many possible ways to define extreme $\widehat{SME}$ values, but one heuristic would be to exclude participants whose $\widehat{SME}$ values are so large that including them would be expected to decrease the effect size of comparisons between group means. A potential method for this is described in Section S8 of the online supplementary materials. However, excluding participants can potentially bias the results of a study, so extensive research would be necessary before deciding on a specific exclusion rule, and the rule for a given study should be determined before the data are seen.

The SME could also be used to identify noisy electrode sites within individual participants, which could then be interpolated. For example, one could identify channels in which the $\widehat{SME}$ value is more than two standard deviations away from the mean $\widehat{SME}$ value for that participant and interpolate the data from those channels. Again, substantial research would be needed before choosing a specific rule.

Formal analyses of these heuristics is beyond the scope of this paper but would be a useful direction for future research.

8. SUMMARY AND POTENTIAL USES

A standardized and widely reported metric of data quality is long overdue in ERP research. The SME metric developed in the present paper meets three key criteria for a metric of data quality: a) it can quantify the data quality in individual participants; b) it reflects the quality of the actual amplitude or latency score being used as the dependent variable in a given experiment; and c) it quantifies the precision of that score. The SME has an additional virtue as well, namely that it can be aggregated across participants in a manner that makes it possible to quantify the contribution of measurement error to the overall observed variability across participants (much like traditional psychometric reliability measures). This makes it possible to estimate the reduction in effect size and statistical power being produced by the measurement error and to predict how much the effect size and statistical power would change as a result of a change in the number of trials or the single-trial noise level.

In some ways, there is nothing new about the SME. We have simply taken the widely used concept of the standard error of measurement and applied it to ERP amplitude and latency scores obtained from single-participant averaged ERP waveforms. However, we know of no previous work proposing that the standard error of measurement should be used as a general metric for ERP data quality, and we know of no previous work showing how single-participant standard errors can be aggregated so that they can be directly related to effect sizes and statistical power in ERP research.

The SME has a large number of potential uses. You could use it to determine whether your data quality has increased or decreased when you modify a data analysis step or experimental design feature. It could indicate that a technical problem has arisen that is degrading the data quality (e.g., degraded electrodes, a poorly trained research assistant). You could use it to determine whether a given participant's data are too noisy to be included in the analyses or whether a channel is so noisy that it should be replaced with interpolated values (but see section S8 in the online supplementary materials for some cautions). If a published paper or submitted manuscript provided SME values, you could use these values to objectively assess whether the data are unacceptably noisy.

Of course, you should not focus solely on the precision of your measures: changing the recording and analysis methods in a manner that creates bias, reduces the validity of the scores, or decreases the true differences in scores between conditions would not usually be a good idea even if these changes reduced the SME. For example, if you simply bridged all of your electrodes to the reference electrode to "flatline" the data on every trial, you would

have a near-zero SME but meaningless data. Similarly, extensive filtering can reduce trial-to-trial variation and therefore decrease the SME, but it can also distort onset times and create artifactual peaks in the ERP waveforms (Acunzo, MacKenzie, & van Rossum, 2012; Tanner, Morgan-Short, & Luck, 2015; Yael, Vecht, & Bar-Gad, 2018). Thus, although a lower SME is better when all else is equal, all else is not always equal, and it will be important not to overemphasize the SME and disregard the many other factors that are important in drawing valid scientific inferences.

Currently, most claims about data quality are largely anecdotal and subjective, making it is difficult to know which ERP recording and analysis methods lead to the cleanest data. However, if ERP papers regularly reported RMS($\widehat{SME}$) values, it would be possible to systematically and objectively compare data quality across experimental paradigms, across EEG recording systems, across signal processing methods, and across data analysis procedures. Researchers could then adopt whatever recording and analysis methods have been shown to produce the best data quality. This would, in turn, lead to increases in statistical power and replicability across the field. However, as discussed earlier, it is important to ensure that these methods for optimizing data quality do not have unintended side effects, such as introducing biases. Also, there are likely to be many factors that differ across studies that could potentially be responsible for differences in RMS($\widehat{SME}$) values between any two studies (e.g., data collection procedures, experimental paradigm details, number of trials, and signal processing steps), making it difficult to ascertain the source of the differences in data quality. However, as more and more studies report SME values, it should be possible to isolate the key factors. Simulation studies will also be valuable for systematically assessing the factors that impact SME; because ground truth is known in simulations, they can also assess whether a given method for reducing SME also introduces biases (see Kiesel, Miller, Jolicoeur, & Brisson, 2008 for an excellent example of this approach).

We therefore recommend that the field move in the direction of consistently reporting RMS($\widehat{SME}$) in published studies[18]. To facilitate this, ERPLAB Toolbox (beginning with version 8) makes it trivially easy to calculate the analytic $\widehat{SME}$ for time-window mean amplitude scores. When you are preparing to compute the averaged ERP waveforms for a given participant, you merely indicate one or more time windows that will be used for scoring the mean amplitude (e.g., 300–500 ms). Default time windows are also provided. ERPLAB will then calculate the analytic $\widehat{SME}$ for each condition during the averaging process (at each electrode site or cluster of electrode sites) by applying Equation 1 to time-window mean amplitude scores obtained from the single-trial EEG epochs.

ERPLAB also includes functions that make it straightforward to calculate the $\widehat{SME}$ for other measures using bootstrapping. Moreover, when you make a grand average across participants, ERPLAB will take the single-participant $\widehat{SME}$ values and compute RMS($\widehat{SME}$) across participants. This gives you everything you need to obtain the $\widehat{SME}$ values. Section

[18]It is certainly possible that even better metrics of data quality will be developed in the future, at which point the field could switch to those new metrics. However, any new metrics must satisfy the three criteria listed in Section 1.1 to be considered a replacement for the SME.

S9 of the online supplementary materials provides recommendations for how these values should be reported for several common ERP experimental designs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank the UC-Davis EEG group for helpful discussions of the standardized measurement error approach, and Kara Federmeier, Dr. Jakub Szewczyk, and an anonymous reviewer for helpful comments on the manuscript.

Funding Information

National Institute of Mental Health grant R01MH087450

References

Acunzo DJ, MacKenzie G, & van Rossum MCW (2012). Systematic biases in early ERP and ERF components as a result of high-pass filtering. Journal of Neuroscience Methods, 209(1), 212–218. 10.1016/j.jneumeth.2012.06.011 [PubMed: 22743800]

Baker DH, Vilidaite G, Lygo FA, Smith AK, Flack TR, Gouws AD, & Andrews TJ (2020). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. ArXiv:1902.06122 [q-Bio, Stat] http://arxiv.org/abs/1902.06122

Balazs A (2008). International Vocabulary of Metrology–Basic and General Concepts and Associated Terms. Joint Committee for Guides in Metrology, 20–21. 10.1515/ci.2008.30.6.21

Boudewyn MA, Luck SJ, Farrens JL, & Kappenman ES (2018). How Many Trials Does It Take to Get a Significant ERP Effect? It Depends. Psychophysiology, 55, e13049. 10.1111/psyp.13049 [PubMed: 29266241]

Brandmaier AM, Wenger E, Bodammer NC, Kühn S, Raz N, & Lindenberger U (2018). Assessing reliability in neuroimaging research through intra-class effect decomposition (ICED). ELife, 7. 10.7554/eLife.35718

Bürki A, Elbuy S, Madec S, & Vasishth S (2020). What did we learn from forty years of research on semantic interference? A Bayesian meta-analysis. Journal of Memory and Language, 114, 104125. 10.1016/j.jml.2020.104125

Bürki A, Frossard J, & Renaud O (2018). Accounting for stimulus and participant effects in event-related potential analyses to increase the replicability of studies. Journal of Neuroscience Methods, 309, 218–227. 10.1016/j.jneumeth.2018.09.016 [PubMed: 30232038]

Chernick MR (2011). Bootstrap Methods: A Guide for Practitioners and Researchers. John Wiley & Sons.

Clayson PE, Baldwin SA, & Larson MJ (2013). How does noise affect amplitude and latency measurement of event-related potentials (ERPs)? A methodological critique and simulation study. Psychophysiology, 50, 174–186. 10.1111/psyp.12001 [PubMed: 23216521]

Clayson PE, Brush CJ, & Hajcak G (2020). Data Quality and Reliability Metrics for Event-Related Potentials (ERPs): The Utility of Subject-Level Reliability [Preprint]. PsyArXiv. 10.31234/osf.io/ja6bw

Ioannidis JP (2005). Why most published research findings are false. PLoS Medicine, 2(8), e124. 10.1371/journal.pmed.0020124 [PubMed: 16060722]

Kiesel A, Miller J, Jolicoeur P, & Brisson B (2008). Measurement of ERP latency differences: A comparison of single-participant and jackknife-based scoring methods. Psychophysiology, 45, 250–274. 10.1111/j.1469-8986.2007.00618.x [PubMed: 17995913]

Lopez-Calderon J, & Luck SJ (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. Frontiers in Human Neuroscience, 8, 213. 10.3389/fnhum.2014.00213 [PubMed: 24782741]

Luck SJ (2012). Electrophysiological correlates of the focusing of attention within complex visual scenes: N2pc and related ERP components. In Luck SJ & Kappenman ES (Eds.), The Oxford Handbook of ERP Components (pp. 329–360). Oxford University Press.

Luck SJ (2014). An Introduction to the Event-Related Potential Technique, Second Edition. MIT Press.

Olvet DM, & Hajcak G (2009). Reliability of error-related brain activity. Brain Research, 1284, 89–99. [PubMed: 19501071]

Picton TW (2011). Human Auditory Evoked Potentials. Plural Publishing.

Smulders FTY, & Miller JO (2012). The Lateralized Readiness Potential. In Luck SJ & Kappenman ES (Eds.), The Oxford Handbook of Event-Related Potential Components (pp. 209–229). Oxford University Press.

Tanner D, Morgan-Short K, & Luck SJ (2015). How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. Psychophysiology, 52, 997–1009. 10.1111/psyp.12437 [PubMed: 25903295]

Thigpen NN, Kappenman ES, & Keil A (2017). Assessing the internal consistency of the event-related potential: An example analysis. Psychophysiology, 54(1), 123–138. 10.1111/psyp.12629 [PubMed: 28000264]

Vasishth S, Mertzen D, Jäger LA, & Gelman A (2018). The statistical significance filter leads to overoptimistic expectations of replicability. Journal of Memory and Language, 103, 151–175. 10.1016/j.jml.2018.07.004

Vazire S (2018). Implications of the credibility revolution for productivity, creativity, and progress. Perspectives on Psychological Science, 13(4), 411–417. 10.1177/1745691617751884 [PubMed: 29961410]

Volpert-Esmond HI, Merkle EC, Levsen MP, Ito TA, & Bartholow BD (2018). Using trial-level data and multilevel modeling to investigate within-task change in event-related potentials. Psychophysiology, 55(5), e13044. 10.1111/psyp.13044 [PubMed: 29226966]

Winsler K, Midgley KJ, Grainger J, & Holcomb PJ (2018). An electrophysiological megastudy of spoken word recognition. Language, Cognition and Neuroscience, 33(8), 1063–1082. 10.1080/23273798.2018.1455985

Wong PK, & Bickford RG (1980). Brain stem auditory evoked potentials: The use of noise estimate. Electroencephalography and Clinical Neurophysiology, 50(1–2), 25–34. [PubMed: 6159189]

Yael D, Vecht JJ, & Bar-Gad I (2018). Filter Based Phase Shifts Distort Neuronal Timing Information. Eneuro. 10.1523/eneuro.0261-17.2018

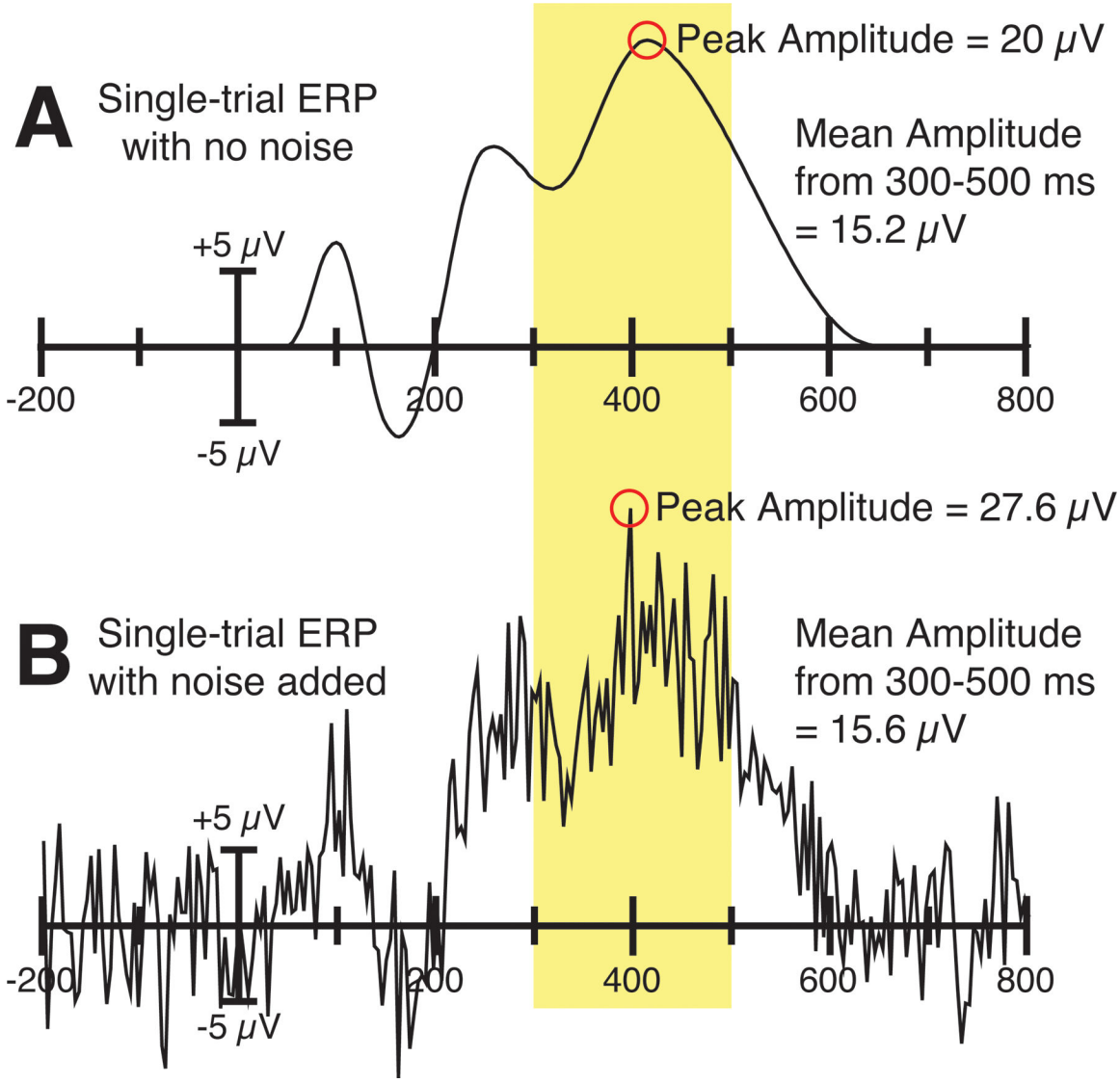

Figure 1.

Example ERP waveform without noise (A) and with substantial high-frequency noise (B). High-frequency noise adds significant variability to measurements of peak voltage (indicated by the red circles). However, it has relatively little effect on time-window mean amplitude measures (e.g., the mean voltage from 300–500 ms).

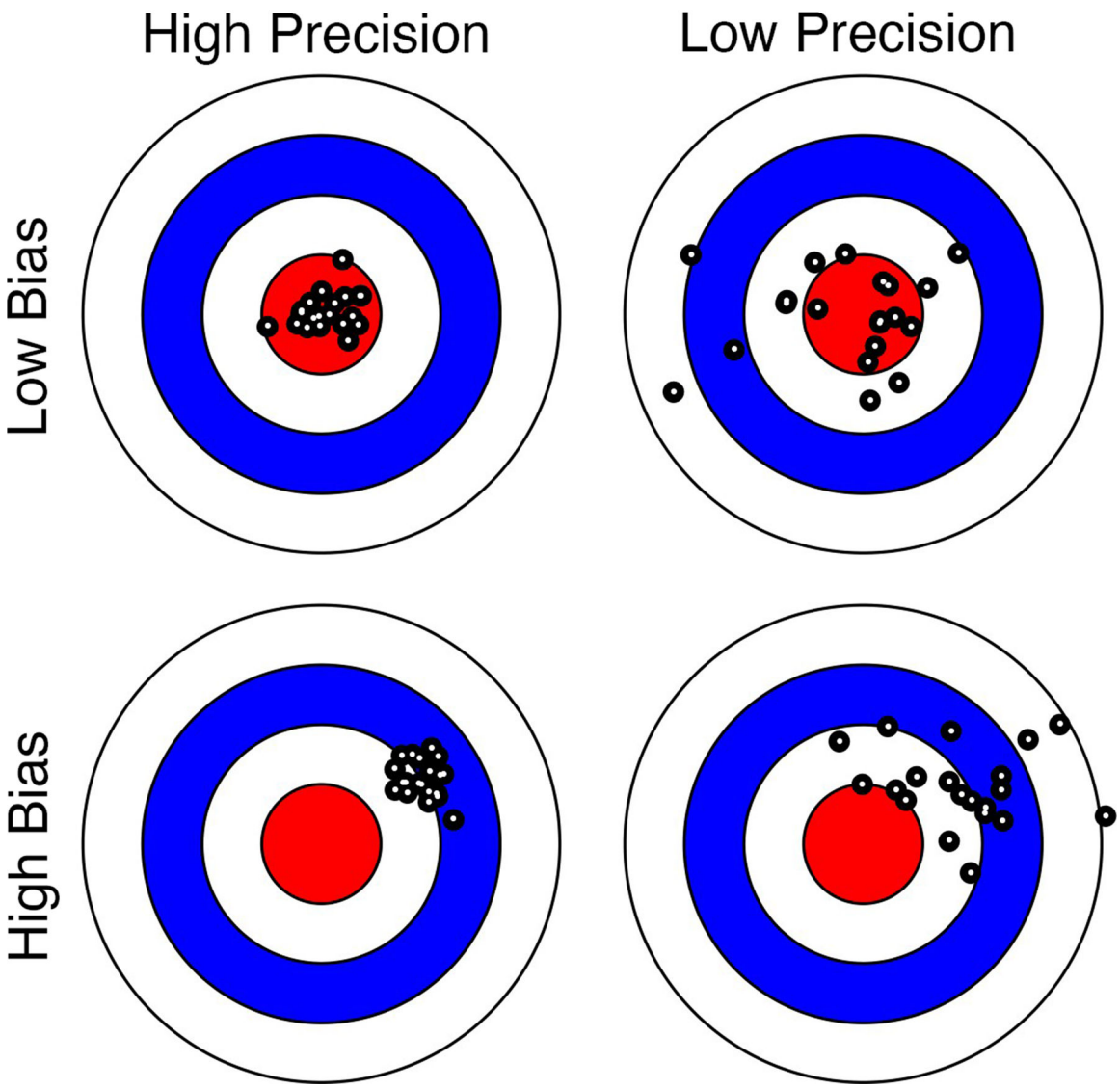

Figure 2.

Graphical depiction of measurement error and the concepts of precision and bias. The center of the bullseye represents the true value that we are attempting to measure (e.g., the true P3 peak latency for a given participant, which is a theoretical quantity rather than an empirically determined value). Each “bullet hole” represents a single attempt to measure that value (e.g., the P3 peak latency observed in an averaged ERP waveform from that participant in a single recording session). A measure of an underlying value is precise to the extent that the values are similar across repeated attempts to measure the value (e.g., similar P3 latency values in averaged ERP waveforms recorded from the same participant in multiple sessions). A metric of precision therefore indicates the spread of values that would be expected across multiple measurement attempts. By contrast, bias reflects the extent to which the average across values is near the true value. A measurement procedure can be biased or unbiased independently of whether it is precise or imprecise. The present measure of data quality focuses solely on precision. Adapted with minor formatting changes from Brandmaier et al.

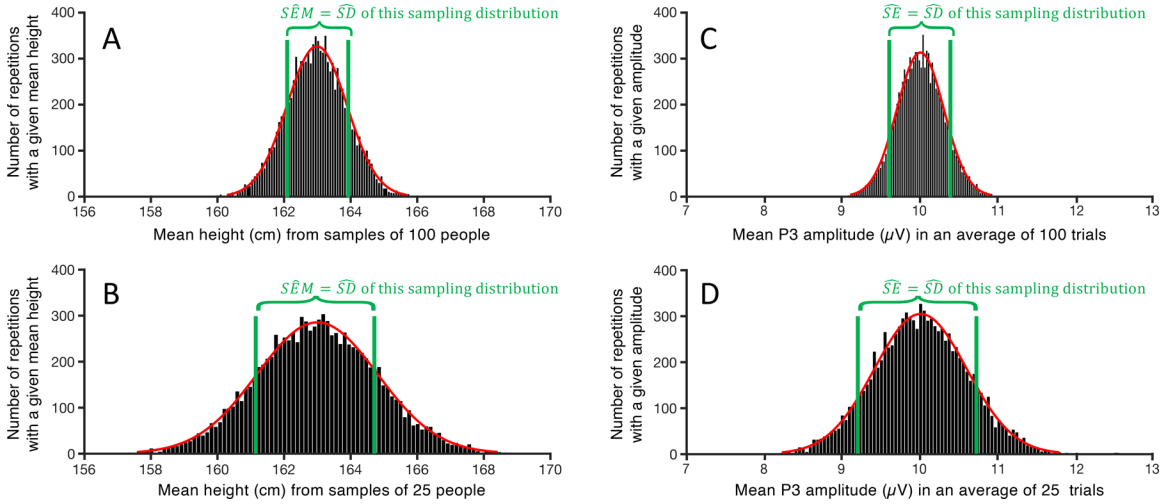

Figure 3.

Empirical approach to estimating the standard error of the mean (SEM) for the height of US women (A and B) and the peak amplitude of the P3 wave in the averaged ERP from a single subject (C and D). (A) In this example, mean height is obtained from a sample of 100 individuals in a given study. The study is repeated 10,000 times, and the frequency distribution of these 10,000 mean height values is shown (which is the sampling distribution of the mean height). Most of the sample means are close to the population mean (approximately 163 cm), but there is some variability. (B) Same as (A) but with samples of 25 people instead of 100 people. These means are more variable than those obtained with samples of 100 people. (C) Extension of the same principle to the P3 peak amplitude of a single participant, measured from an averaged ERP waveform based on 100 trials. The session is repeated 10,000 times, and the P3 peak amplitude is measured from the averaged ERP waveform for each repetition. Most of the measured amplitude scores are near the true score (10 μV), but there is some variability. (D) Same as (C) but with an ERP waveform created by averaging together 25 trials instead of 100 trials. The resulting P3 amplitude scores are more variable than those obtained with averages of 100 trials. In each of these four examples, the standard error ((SEM) ^) is estimated by taking the standard deviation of the 10,000 values in the sampling distribution.

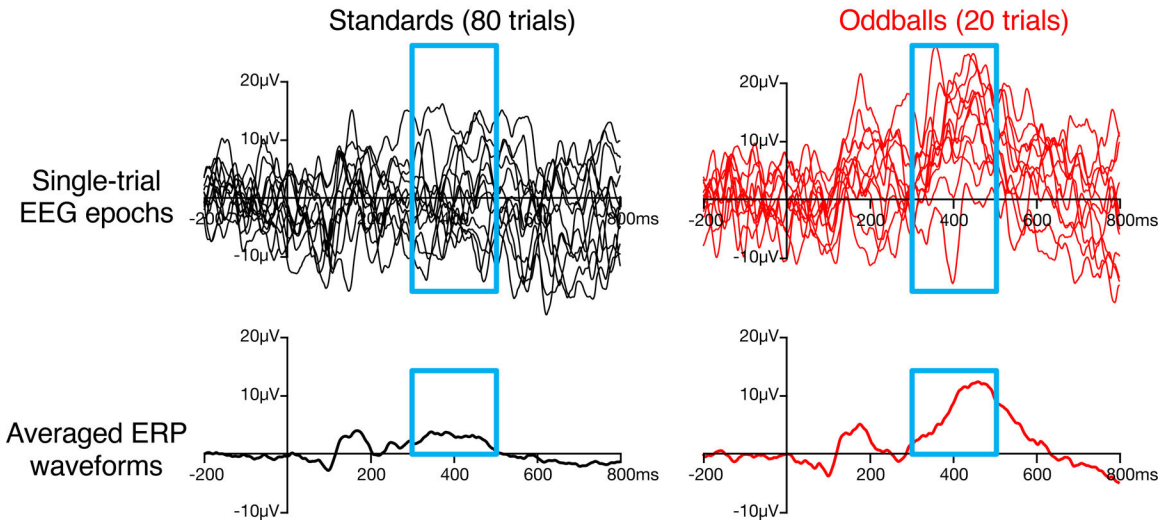

Figure 4.

Example of the single-trial EEG epochs and corresponding averaged ERP waveforms from a single participant in an oddball experiment. Only a subset of the single-trial EEG epochs are shown. The blue boxes indicate the period used to compute the time-window mean amplitude (300–500 ms). This score is ordinarily obtained from the averaged ERP waveforms. The SME for this score is computed by obtaining the score from the single-trial EEG epochs and applying Equation 1.

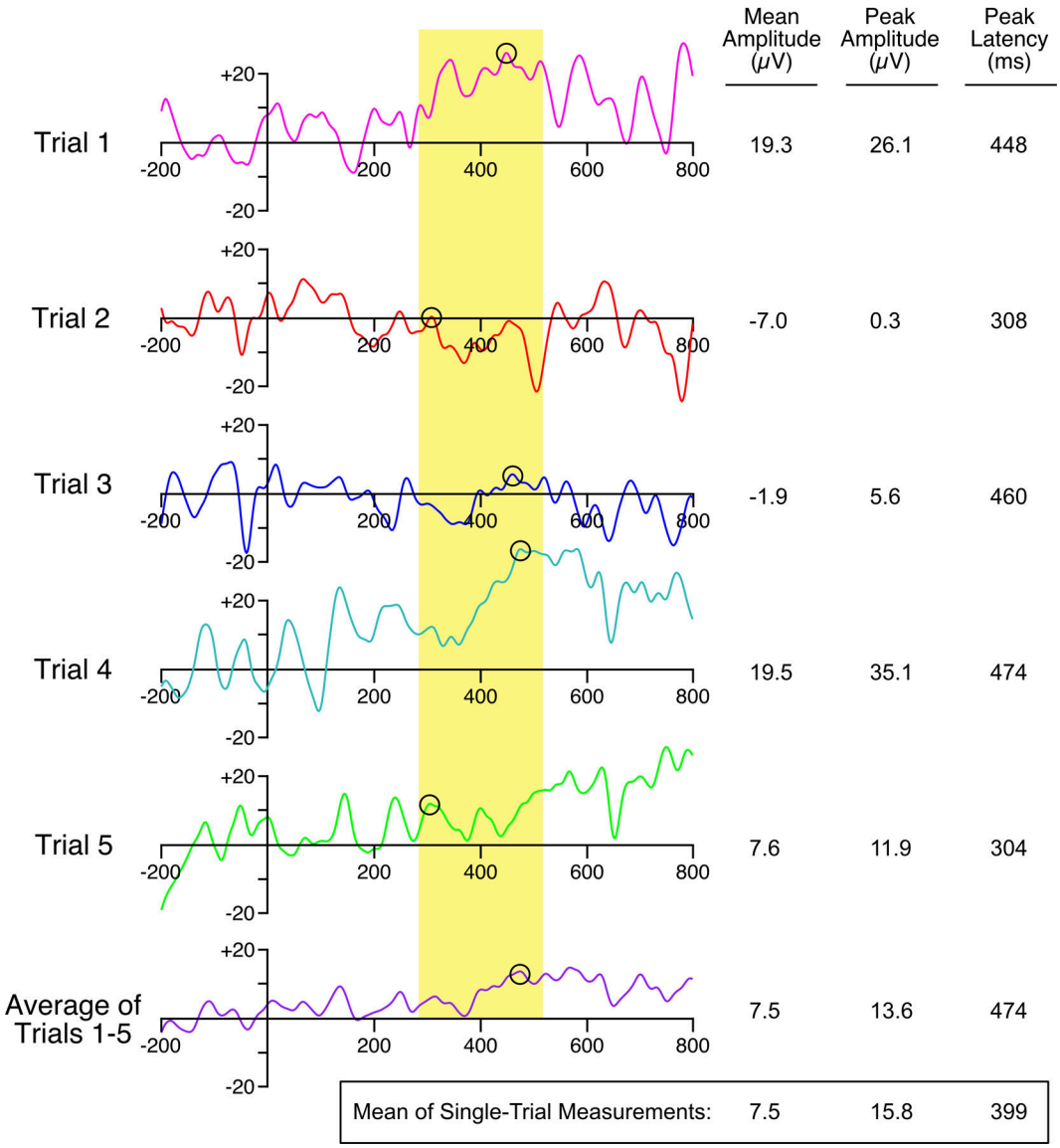

Figure 5.

Example of scoring the time-window mean amplitude or the peak amplitude from single-trial ERP epochs or from an averaged ERP waveform. Five single-trial EEG epochs are shown, along with the average of these 5 epochs. The time-window mean amplitude and peak amplitude measures were obtained (using a time window of 300–500 ms) from each single-trial epoch and also from the averaged waveform. The mean of the time-window mean amplitudes from the five individual trials is identical to the time-window mean amplitude measured from the averaged waveform. However, the mean of the peak amplitudes of the five EEG epochs is not the same as the peak amplitude of the averaged waveform.

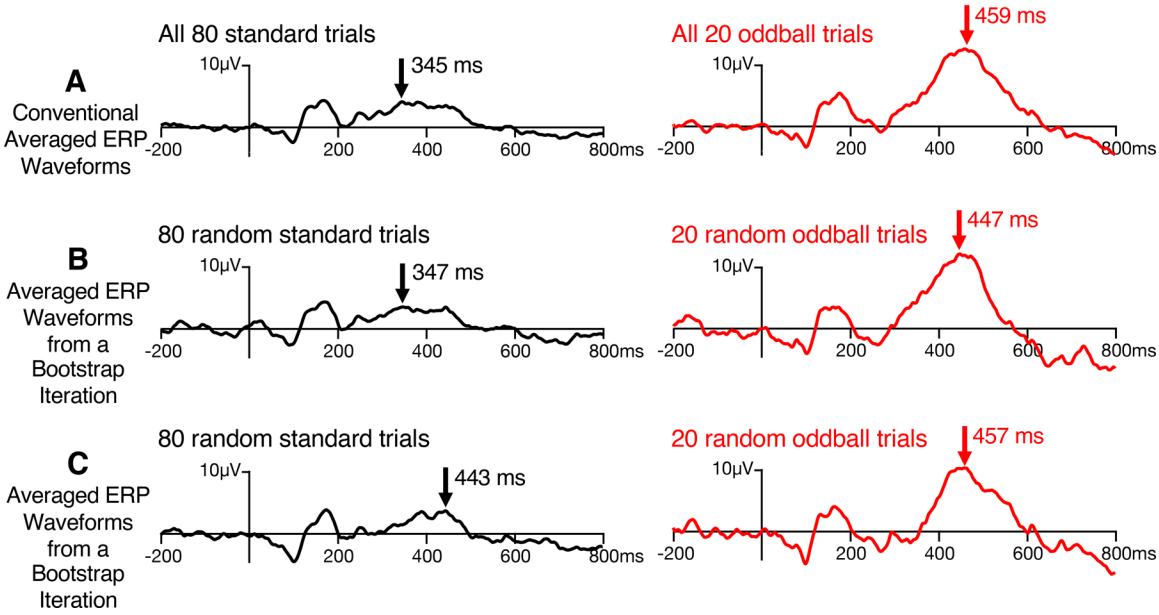

Figure 6.

Example of measuring P3 peak latency using bootstrapped ERP waveforms from a single participant. There were 80 standard trials and 20 oddball trials, and (A) shows the averages of these 80 standards and 20 oddballs. Averages created from two bootstrap iterations are shown in (B) and (C). For each bootstrap iteration, 80 trials were selected at random with replacement from the set of 80 standards, and 20 trials were selected at random with replacement from the set of 20 oddballs. The peak latency is shown for each averaged ERP waveform.

Table 1.

Amplitude and latency scores along with the corresponding analytic standardized measurement error ($\widehat{aSME}$) and bootstrapped standardized measurement error ($\widehat{bSME}$) values for each participant. Also provided are the mean across participants (sample mean), the standard deviation across participants (sample $\widehat{SD}$), and the root mean square of the standardized measurement error values (RMS($\widehat{SME}$)).

Subject ID	Time-Window Mean Amplitude (Standards)			Time-Window Mean Amplitude (Oddballs)			Peak Amplitude (Standards)		Peak Amplitude (Oddballs)		Peak Latency (Standards)		Peak Latency (Oddballs)	
	Score (uV)	$\widehat{aSME}$	$\widehat{bSME}$	Score (uV)	$\widehat{aSME}$	$\widehat{bSME}$	Score (uV)	$\widehat{bSME}$	Score (uV)	$\widehat{bSME}$	Score (ms)	$\widehat{bSME}$	Score (ms)	$\widehat{bSME}$
1	6.52	2.08	2.07	6.55	4.75	4.70	10.36	1.83	11.20	5.48	325.20	20.28	489.26	23.75
2	2.01	0.82	0.82	6.09	1.59	1.54	4.59	0.94	10.67	1.80	303.71	41.89	491.21	42.27
3	6.42	0.91	0.91	7.72	1.64	1.61	7.89	1.04	10.37	2.02	396.48	35.76	422.85	55.89
4	0.77	0.81	0.81	7.79	1.42	1.37	2.08	0.86	10.04	1.55	308.59	61.51	423.83	23.46
5	6.81	2.12	2.11	7.09	2.89	2.84	8.11	2.33	11.84	2.99	481.45	41.89	444.34	9.07
6	4.81	0.70	0.69	12.48	1.12	1.09	7.49	0.91	18.00	1.68	341.80	39.92	405.27	13.79
7	−0.11	0.75	0.74	0.43	1.53	1.47	1.24	0.79	2.59	1.77	390.63	29.63	491.21	50.39
8	4.70	1.00	1.00	2.30	2.38	2.29	7.54	1.13	9.08	2.37	426.76	18.86	438.48	6.70
9	5.33	1.07	1.08	8.85	1.71	1.67	7.05	1.31	12.81	1.63	400.39	24.10	441.41	24.83
10	7.36	0.77	0.76	14.04	1.51	1.46	9.20	0.95	17.09	1.68	424.80	39.26	422.85	42.34
11	4.82	0.51	0.50	4.66	1.15	1.12	6.37	0.60	6.00	1.35	392.58	22.33	391.60	39.27
12	3.09	0.56	0.55	8.18	1.28	1.25	4.11	0.63	12.69	1.55	344.73	35.60	458.98	16.79
Sample Mean	4.38	1.01	1.00	7.18	1.91	1.87	6.34	1.11	11.03	2.15	378.09	34.25	443.44	29.05
Sample $\widehat{SD}$	2.44			3.78			2.78		4.20		53.96		33.45	
RMS($\widehat{SME}$)		1.13	1.13		2.15	2.11		1.21		2.41		36.17		33.06